



The Open
University

Faculty of Science, Technology, Engineering and Mathematics
M249 Practical modern statistics

M249

TMA02

2017J

Cut-off date 31 January 2018

You can submit each TMA either by post or electronically using the online TMA/EMA service. Please read the guidance under the 'Assessment' tab of the M249 website.

Each TMA is marked out of 100. The marks allocated to each part of each question are indicated in brackets in the margin. Your overall score for each TMA will be the sum of your marks for all questions in that TMA.

Questions 1 to 3 below, on Book 2 *Time series*, form tutor-marked assignment M249 02. Question 1 (on Part I of Book 2) is marked out of 33. Question 2 (on Part II) is marked out of 35. Question 3 (on Part III) is marked out of 32.

Question 1 – 33 marks

This question is intended to assess your understanding of the use and interpretation of time series plots, moving averages and methods for estimating the components of a time series. You should be able to answer this question after working through Part I of Book 2.

In this question you will be required to supply SPSS output for parts (a)(i), (a)(ii), (b)(i), (b)(ii), (b)(iii) and (b)(iv) only, though you will be expected to use SPSS to answer the rest of the question. In parts (a)(i) and (a)(ii), you will need to edit default SPSS plots, and you should include only the edited plots in your work. All SPSS output should be included in the body of your work at the relevant point, and you should include only what is relevant to the question and your answer.

- (a) The file **arctic.sav** contains data on the extent of sea ice in the Chukchi Sea, which lies between Siberia and Alaska, north of the Bering Strait. Measurements (in thousand km²) are taken in August of each year.
- (i) The variable **ice** contains the differences, for the years 1900–2000, between the extent of sea ice and a fixed baseline value. Obtain a time plot of **ice** with diagonal time axis labels at ten-year intervals (1900, 1910, . . . , 2000). Include a copy of your final plot in your answer. Describe the main features of the time plot. [4]
- (ii) Obtain three time series comprising centred moving averages of **ice** of orders 3, 9 and 29. Obtain separate time plots for each of these series, with the same vertical scale as the time plot that you obtained in part (a)(i). Include copies of these three plots in your answer. [4]
- (iii) State which of the three moving averages that you obtained in part (a)(ii) provides the best estimate of the trend component. Justify your choice. [3]
- (iv) Briefly describe the trend in the extent of sea ice in the Chukchi Sea. [3]

- (b) The file **yoghurt.sav** contains data on the monthly production of frozen yoghurt in the USA between January 1989 and December 2009.
- (i) The variable **production** contains the monthly production in thousands of pounds in weight (lb). Obtain a time plot of **production**, and include a copy of your plot in your answer. Comment on the main features of the time series. Explain why an additive model is not appropriate to model **production**. [4]
 - (ii) The variable **logproduction** contains the natural logarithm of **production**. Obtain a time plot of **logproduction**, and include a copy of your plot in your answer. Give two reasons why it may be reasonable to represent **logproduction** using an additive model. [3]
 - (iii) Use SPSS to obtain the seasonal factors for **logproduction**, and include the resulting table of seasonal factors in your answer. Describe the seasonal variation in the production of frozen yoghurt over this period. Tidy up the data file, retaining the seasonally adjusted series, which you should rename **adjusted**, and obtain a time plot of **adjusted**. Include a copy of this time plot in your answer. [5]
 - (iv) Using a centred moving average of order 11, obtain estimates of the trend component and of the irregular component of **logproduction**. Explain clearly the methods that you use. Obtain time plots of the trend component and of the irregular component, and include copies of these time plots in your answer. [4]
 - (v) Briefly describe the estimated components of **logproduction**, and comment on their relative importance. [3]

Question 2 – 35 marks

This question is intended to assess your understanding of the selection, use and interpretation of exponential smoothing methods for time series. You should be able to answer this question after working through Part II of Book 2.

In this question you will be required to supply SPSS output for parts (a)(i), (a)(iv), (a)(v), (b)(i) and (b)(iv) only, though you will be expected to use SPSS to answer the rest of the question. All SPSS output should be included in the body of your work at the relevant point, and you should include only what is relevant to the question and your answer.

- (a) The SPSS data file **oilprices1.sav** contains the prices (in US dollars) per barrel of Brent Crude oil on the first trading day of each week from 30 July 2012 to 28 July 2014. (Note that a time variable has not been defined in this SPSS data file because SPSS cannot define dates for weekly data.)
- (i) Obtain a time plot of **price**, and include a copy of the plot in your answer. Briefly describe the time plot. [3]
 - (ii) Using simple exponential smoothing, obtain the optimal value of the smoothing parameter. Interpret the value of this parameter in terms of the relative weight given to recent and past observations. Obtain the corresponding *RMSE*, and hence calculate the value of the *SSE*. [4]

- (iii) The 1-step ahead forecast for the last value of **price** is $\hat{x}_{105} = 105.66$. Without using SPSS, obtain the 1-step ahead forecast for the next value of price, $\hat{x}_{106}$ (corresponding to 4 August 2014), and calculate an approximate 95% prediction interval for it. Show your working. [4]
- (iv) Obtain the correlogram of the forecast errors for lags 1 to 20, and include a copy of the plot in your answer. Carry out the Ljung–Box test for zero autocorrelation at lags 1 to 20. What do you conclude about the presence or otherwise of non-zero autocorrelations between the forecast errors at lags 1 to 20? [4]
- (v) Obtain a time plot and a histogram (with overlaid normal curve) of the forecast errors. Include copies of your plots in your answer. Are the forecast errors normally distributed with mean zero and constant variance? Explain your answer. [5]
- (vi) Discuss whether the assumptions required for calculating the prediction interval in part (a)(iii) seem valid. In answering this, you should refer to your answers to previous parts of this question. [2]
- (b) This part of the question relates to data on daily prices per barrel of Brent Crude oil (in US dollars) between 4 August 2014 and 13 February 2015. The data are in the SPSS file **oilprices2.sav**.
- (i) Obtain a time plot of the variable **price**, and include a copy of the plot in your answer. Discuss which type of exponential smoothing method is appropriate for these data. [3]
- (ii) It is decided to use Holt’s exponential smoothing method. Explain why the following starting values are reasonable: 103.63 for the level and -0.81 for the trend. [2]
- (iii) Obtain and report the optimal values of the parameters for forecasting the values of price using Holt’s method. Interpret the values of the two parameters in terms of the relative weight given to recent and past observations. What, in particular, does the obtained value of γ imply about this time series? [4]
- (iv) Obtain forecasts of price for the next 30 days (days 126 to 155). Display observed values for days 1 to 125, and forecasted values for days 1 to 155, on a multiple time plot. Include a copy of the plot in your answer. Briefly describe the predictions for the next 30 days (days 126 to 155), and comment on their reliability. [4]

Question 3 – 32 marks

This question is intended to assess your understanding of the selection, use and interpretation of ARIMA models for time series. You should be able to answer this question after working through Part III of Book 2.

In this question you will be required to supply SPSS output for parts (a)(i), (a)(ii), (b)(iii) and (b)(iv) only, though you will be expected to use SPSS to answer the rest of the question. All SPSS output should be included in the body of your work at the relevant point, and you should include only what is relevant to the question and your answer.

This question relates to annual data on wheat prices between 1264 and 1996. To make the prices comparable, they have been expressed in 1996 values. The data are in **allwheat.sav**. The variable **logprice** contains the natural logarithms of the wheat prices.

- (a) In this part of the question you will investigate suitable ARIMA models to fit to the time series **logprice**.
- (i) Show that the order of differencing d required to make the time series **logprice** stationary is $d = 1$. Explain your reasoning, and illustrate it with appropriate time plots. [6]
 - (ii) Obtain the correlogram and partial correlogram of the first differences of **logprice**, for lags 1 to 20. Include copies of the correlogram and partial correlogram in your answer. Describe the main features of the correlogram and the partial correlogram. [4]
 - (iii) One plausible model for **logprice** is ARIMA(1, 1, 1). Using the plots obtained in part (a)(ii), explain carefully why this model is plausible. [4]
 - (iv) Suggest two further plausible models for **logprice**, where one model has $p = 0$ and the other model has $q = 0$. Using the principle of parsimony, identify one or more ‘best’ candidate models from those that you have identified and that specified in part (a)(iii). Explain carefully the reasons for the choices that you make. [4]

- (b) In this part of the question you will fit the ARIMA(1, 1, 1) model to the time series `logprice` and investigate its adequacy.
- (i) Fit the ARIMA(1, 1, 1) model to `logprice`, and obtain the estimates of the parameters μ , β_1 and θ_1 . Let X_t denote the first differences of the `logprice` values. Write down the model equation for the X_t using the estimated parameter values. [4]
 - (ii) What is the value of the *SSE* associated with the model fitted in part (b)(i)? [2]
 - (iii) Using appropriate plots, check that the forecast errors have mean zero and constant variance, and that they are approximately normally distributed. Include copies of these plots in your answer. [4]
 - (iv) Obtain the correlogram of the forecast errors at lags 1 to 20, and include a copy in your answer. Carry out the Ljung–Box test of the null hypothesis that autocorrelations of the forecast errors at lags 1 to 20 are zero. Are the autocorrelations of the forecast errors at lags 1 to 20 equal to zero? Give one reason for, and one reason against. [3]
 - (v) Based on your answers to parts (b)(iii) and (b)(iv), briefly discuss the adequacy of the ARIMA(1, 1, 1) model. [1]
-